Python and its application on CNV data

BIO392 day3

Ziying Yang

Python

🐍 What is Python?

Python is a high-level, interpreted programming language known for its **simplicity**, **readability**, and **versatility**. It was created by **Guido van Rossum** and first released in **1991**.

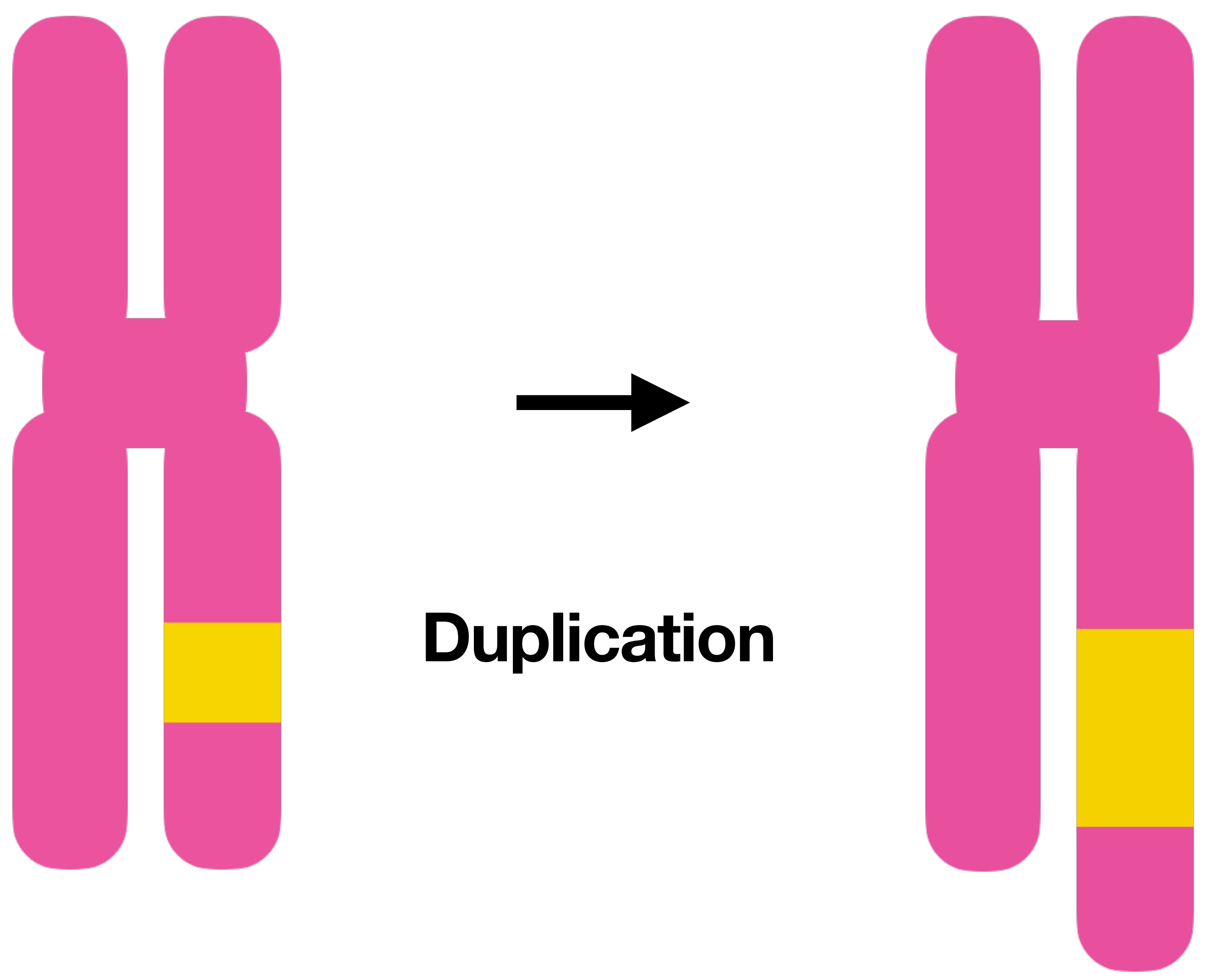
🔷 Why is Python so popular?

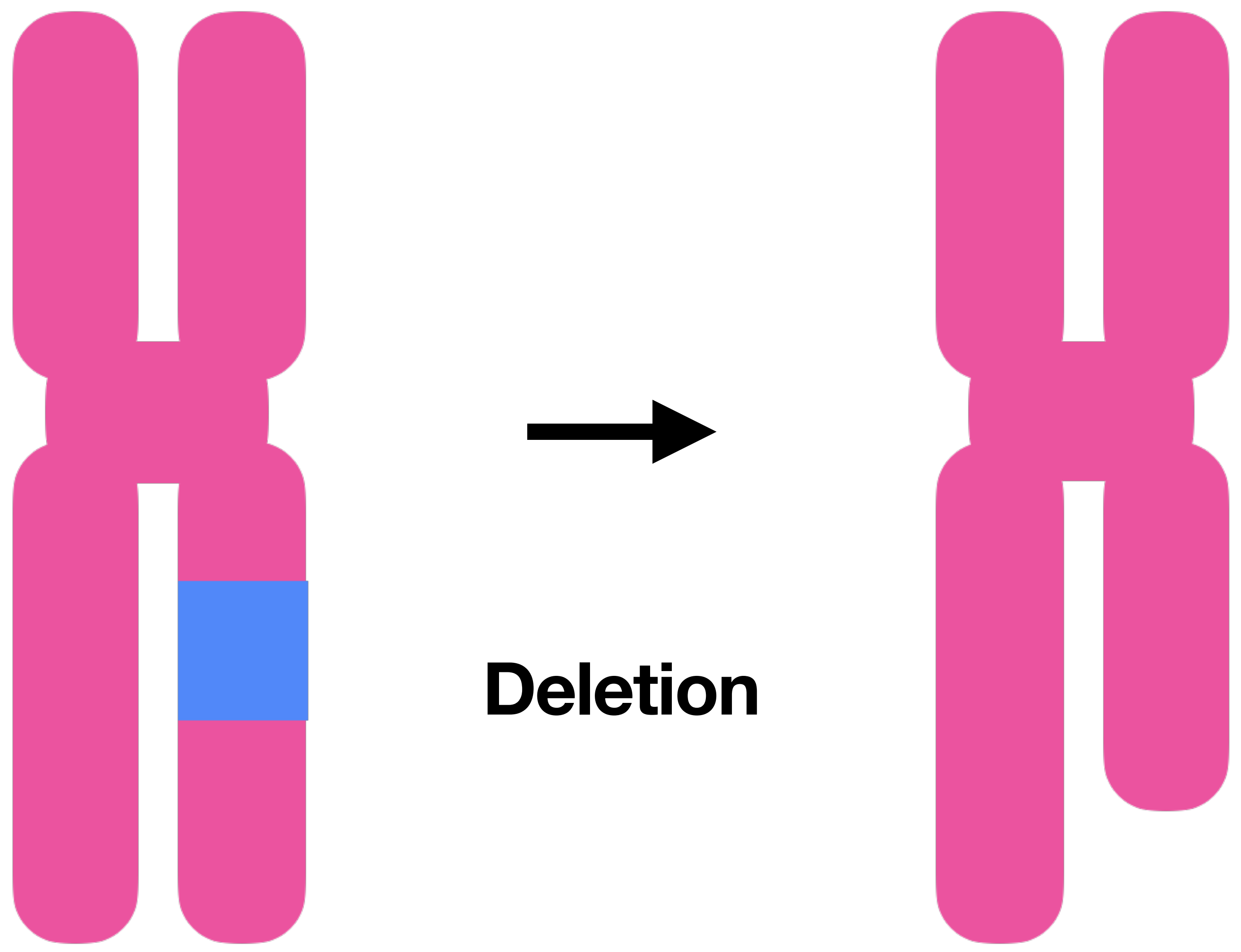
- **Easy to read & write**: Looks almost like English
- **Huge ecosystem**: Tons of libraries for data, web, AI, etc.
- **Cross-platform**: Runs on Linux, Windows, macOS
- **Community support**: Massive open-source backing

🔶 Where is Python used?

- **Data Science & Machine Learning**: pandas, numpy, scikit-learn, tensorflow
- **Web Development**: Django, Flask, FastAPI
- **Automation & Scripting**: os, subprocess, shutil
- **Bioinformatics**: Biopython, pysam, custom pipelines
- **DevOps & Cloud**: boto3, ansible

Copy Number Variant (CNV)

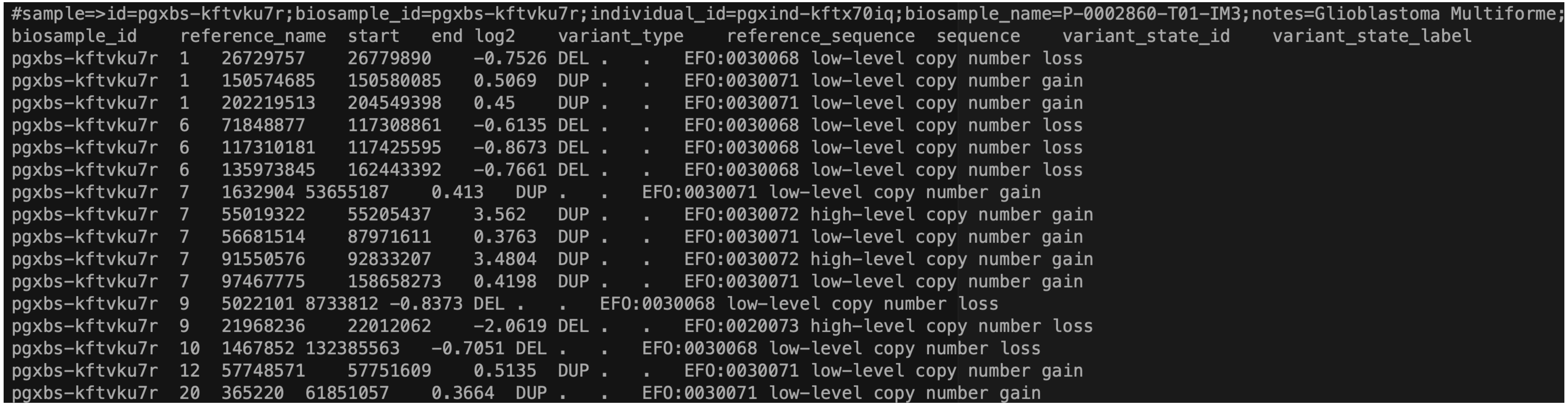




- Intermediate-scale genetic change
- Size: 1kb to multiple megabase
- Additional copies of sequence (duplications) and losses of genetic material (deletions)

Python warm-up on bioinformatics

CNV segmentation data



```
#sample=>id=pgxbs-kftvku7r;biosample_id=pgxbs-kftvku7r;individual_id=pgxind-kftx70iq;biosample_name=P-0002860-T01-IM3;notes=Glioblastoma Multiforme;
biosample_id    reference_name  start   end log2    variant_type    reference_sequence  sequence    variant_state_id    variant_state_label
pgxbs-kftvku7r  1   26729757    26779890    -0.7526 DEL .   .   EFO:0030068 low-level copy number loss
pgxbs-kftvku7r  1   150574685   150580085   0.5069  DUP .   .   EFO:0030071 low-level copy number gain
pgxbs-kftvku7r  1   202219513   204549398   0.45    DUP .   .   EFO:0030071 low-level copy number gain
pgxbs-kftvku7r  6   71848877    117308861   -0.6135 DEL .   .   EFO:0030068 low-level copy number loss
pgxbs-kftvku7r  6   117310181   117425595   -0.8673 DEL .   .   EFO:0030068 low-level copy number loss
pgxbs-kftvku7r  6   135973845   162443392   -0.7661 DEL .   .   EFO:0030068 low-level copy number loss
pgxbs-kftvku7r  7   1632904 53655187    0.413   DUP .   .   EFO:0030071 low-level copy number gain
pgxbs-kftvku7r  7   55019322    55205437    3.562   DUP .   .   EFO:0030072 high-level copy number gain
pgxbs-kftvku7r  7   56681514    87971611    0.3763  DUP .   .   EFO:0030071 low-level copy number gain
pgxbs-kftvku7r  7   91550576    92833207    3.4804  DUP .   .   EFO:0030072 high-level copy number gain
pgxbs-kftvku7r  7   97467775    158658273   0.4198  DUP .   .   EFO:0030071 low-level copy number gain
pgxbs-kftvku7r  9   5022101 8733812 -0.8373 DEL .   .   EFO:0030068 low-level copy number loss
pgxbs-kftvku7r  9   21968236    22012062    -2.0619 DEL .   .   EFO:0020073 high-level copy number loss
pgxbs-kftvku7r  10  1467852 132385563   -0.7051 DEL .   .   EFO:0030068 low-level copy number loss
pgxbs-kftvku7r  12  57748571    57751609    0.5135  DUP .   .   EFO:0030071 low-level copy number gain
pgxbs-kftvku7r  20  365220  61851057    0.3664  DUP .   .   EFO:0030071 low-level copy number gain
```

https://progenetix.org/services/pgxsegvariants?biosample_ids=pgxbs-kftvku7r

Python warm-up

- Data link: *https://progenetix.org/services/pgxsegvariants?biosample_ids=pgxbs-kftvku7r,pgxbs-m3io2hj8,pgxbs-kftvkuvy*
- Check the data first, and write your own script to access and download the data via python (*tips: requests*).
- Transfer the data to dataframe in pycharm (*tips: pandas.dataFrame()*), with proper columns.

Python warm-up

- ***Histplot***: You can start by exploring the data to understand its structure and distribution. For example, you can check the distribution of the ‘reference_name' values using a histogram
- ***Count plot***: Count the number of CNV events per biosample
- ***Heatmap*** of CNV Events: If you want to explore relationships between biosamples and CNV events, you can create a heatmap to visualize the presence or absence of CNV events across biosamples.

https://seaborn.pydata.org/generated/seaborn.histplot.html https://doi.org/10.1093/database/baab043

- What is CNV/CNA?
- How will you describe or introduce progenetix (scale, data source, cancer types and so on)?
- Describe NCIt, ICOD, UBERON codes, and their relationships.
- What are CNV segmentations and CNV frequencies, and how to use them?
- What are APIs and how to use APIs in progenetix?
- How does progenetix visualise CNA profiles?
- What do you think should be improved in progenetix?

Please upload your file to https://github.com/compbiozurich/UZH-BIO392/tree/master/course-results/day3, and name the file as lastname_firstname_paper_reading_day3.md. It will be graded.

https://progenetix.org/

https://docs.github.com/en/get-started/writing-on-github/getting-started-with-writing-and
formatting-on-github/basic-writing-and-formatting-syntax